DOCUMENT RESUME

ED 214 984 TM 820 236

AUTHOR Alderman, Donald L.
TITLE Measurement Error and SAT Score Change.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY College Entrance Examination Board, New York, N.Y.
REPORT NO CEEB-RR-81-9; ETS-RR-81-39
PUB DATE 81
NOTE 21p.; Small print throughout.
AVAILABLE FROM College Board Publication Orders, Box 2815, Princeton, NJ 08541 ($4.00).

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS College Entrance Examinations; *Error of Measurement; *Regression (Statistics); *Scores; Secondary Education; Testing Problems
IDENTIFIERS *Change Scores; *Scholastic Aptitude Test; Test Repeaters

ABSTRACT

This study applies a procedure which yields estimates of true score change on the Scholastic Aptitude Test (SAT) adjusted for regression effects and student self-selection. It is shown that student self-selection in deciding to repeat an admissions test probably involves factors in addition to the measurement error attributable to variations in aspects of test specifications and to variations in responses of test candidates across forms, and that estimated true score change remains nearly constant across initial score levels in contrast to the negative slope of observed score change across initial score levels. (Author)

College Board Report

No. 81-9

Measurement Error and SAT Score Change

Donald L. Alderman

Measurement Error and SAT Score Change

Donald L. Alderman

Educational Testing Service

College Board Report No. 81-9

ETS RR No. 81-39

College Entrance Examination Board, New York, 1981

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publication Orders, Box 2815, Princeton, New Jersey 08541. The price is $4.

CONTENTS

ABSTRACT

Score changes on admissions tests such as the Scholastic Aptitude Test (SAT) can vary widely among individuals repeating the test. To a large extent these score changes probably reflect the influence of errors of measurement as test candidates with low initial scores usually experience score gains upon retesting while test candidates with high initial scores often show score losses. Besides this phenomenon of scores regressing toward the mean upon test repetition, student self-selection may affect score change as test candidates who perceive their initial scores as underestimates of their true abilities decide to repeat the test. This study applies a procedure which yields estimates of true score change on the SAT adjusted for regression effects and student self-selection. It is shown that student self-selection in deciding to repeat an admissions test probably involves factors in addition to the measurement error attributable to variations in aspects of test specifications and to variations in responses of test candidates across forms, and that estimated true score change remains nearly constant across initial score levels in contrast to the negative slope of observed score change across initial score levels.

INTRODUCTION

Each year several hundred thousand students who had taken the Scholastic Aptitude Test (SAT) as high school juniors elect to take the test again as seniors. The distribution of these junior-senior score changes, typically with a mean near 15 points and a standard deviation in the neighborhood of 50 points (Donlon and Angoff, 1971; Educational Testing Service, 1979), includes fairly frequent gains in excess of 65 points as well as regular losses beyond 35 points on the SAT's 200-800 point scale. One-third of the students who take the SAT as juniors and again as seniors will experience score changes greater than 50 points. For every hundred students repeating the test, five students will gain at least 100 points and one student will lose at least 100 points (Educational Testing Service, 1979).

The widespread of such score changes on an admissions test invites misinterpretation. Test candidates unfamiliar with the technical aspects of measurement will find it difficult to understand dramatic score changes in terms of inherent error. It is far simpler for students, parents and counselors to accept all score increases as a reflection of real growth in academic ability or to attribute them to the effectiveness of some intervening event (e.g., a particular course or instructor, a specific program of test preparation). All score decreases, on the other hand, may be viewed as a sign of procedural error on the part of the test publisher or as evidence of further decline in the quality of education. Thus, gains among selected students have been interpreted as indicative of the impact of special test preparation while losses among other selected students have prompted considerable concern about possible implications for school practices (e.g., Kendrick, 1967; Messick, 1980). The extent to which these score changes represent the consequences of measurement error seems unclear but certainly bears on the interpretation of score differences.

Regression Effects and Student Self-Selection

The fallibility of the test as a measurement instrument usually accounts for most of the score difference observed between administrations of a test, as Duggan (1959) suggests is the case for score changes on the SAT. The unreliability and dispersion of SAT score changes follow from the reliability of the test, the test-retest correlation, and the test's standard deviation according to standard formulas such as those given by Lord (1963), McNemar (1958), and Overall and Woodward (1975). For students taking the test in May 1979 as juniors and again in November 1979 as seniors, it can be shown that the reliability of differences in SAT-Verbal scores is only .25 and of differences in SAT-Mathematical scores only .23 (see Alderman, 1981 for estimates of test reliability among repeaters and Tables 1 and 2 below for standard deviations and test-retest correlations). Similarly, the standard deviations of these difference scores are 48 points for the SAT-V and 53 points for the SAT-M. Since there will be approximately 32 percent of the cases at least one standard deviation away from the mean and 5 percent at least two standard deviations away from the mean in any normal distribution, the frequency of SAT score changes greater than 50 points and greater than 100 points conforms to expectations for the test. And the degree of importance attached to the observed score changes for an individual or a selected group of individuals should be consistent with the low reliability of these score changes.

The score fluctuations evident between administrations of the same test largely reflect the effect of scores regressing toward the mean upon the repetition of a test (Lord, 1956, 1958, 1963). A test such as the SAT is most reliable and differentiates best among students near the mean of the distribution of scores. At the extremes of the score scale, however, the test is less reliable as errors of measurement have probably served to depress low scores and to raise high scores to some extent. Consequently, an examinee whose initial score was either very low or very high may experience a score change far greater than the standard error of measurement upon repeating the test. Because the SAT has a standard error of measurement usually near 30 points (e.g., Educational Testing

Service, 1980) and a standard deviation for difference scores close to 50 points, the majority of test candidates repeating the SAT will actually experience score changes, gains and losses, greater than the standard error of measurement. Sizeable score changes are most apparent among students with extreme low scores on an initial administration of the SAT as real growth in ability between test administrations acts to exaggerate further score increments due to regression. For extreme high scores, real growth acts in the opposite direction as it diminishes the magnitude of score decrements due to regression.

The interaction of regression toward the mean upon repetition of a test and real growth between test administrations in the trait reflected by scores on a test can lead to unusual and sometimes counter-intuitive contrasts between observed score changes found through simple differences and true score changes free of the influences of errors of measurement. For test candidates with high initial scores the observed score change is apt to be negative while the true score change is probably positive. Also, an observed loss at one score level may reflect greater growth in ability than an observed gain at another score level as when an individual with a high initial score loses points and an individual with a low initial score gains points on a subsequent test. Empirical evidence of scores regressing toward the mean on multiple administrations of an admissions test appears in Rock and Werts (1980), and a thorough discussion of possible theoretical patterns of score change due to regression across multiple test administrations appears in Nesselroade, Stigler, and Baltes (1980).

Besides regression effects as a component of score change upon repetition of a test, there may be student self-selection in deciding to repeat a test prompted by negative errors of measurement on the initial test administration (e.g., Jacobs, 1966). Alderman (1981) demonstrates that the test scores expected on the basis of student performance on separate equating sections and student background variables such as high school rank and years of mathematics study overpredict the initial verbal and mathematical scores obtained by students who later repeat the SAT. Students electing to repeat an admissions test apparently do so in part because they perceive their initial scores as underestimates of their true abilities. This implies a nonzero, negative sum of errors of measurement on repeaters' initial test scores which would make the observed difference in mean scores between test administrations an overestimate of true score change (e.g., Lord, 1956) and preclude the application of existing models for measuring change (e.g., Lord, 1963).

The purpose of this study is to estimate score change while correcting for regression effects and student self-selection. The method adopted for estimating score change among students taking the SAT as juniors and repeating the test as seniors depends on a concurrent verbal or mathematical score available through shorter but otherwise parallel equating sections. Each administration of the SAT includes a separate experimental section for equating test forms or for pretesting items which does not enter into the operational scores reported back to test candidates and thus should not be a factor in student self-selection in deciding to repeat the test. Using scores on equating sections and adjusting for regression toward the mean should lead to estimates of score change more stable across initial score levels rather than a pronounced negative correlation between observed score change and initial score (e.g., Linn and Slinde, 1977). It should also lead to a lower estimate of mean score change if negative errors of measurement on an initial test administration prompt students to repeat a test. Such findings would indicate the influence of measurement error on observed score change and clarify the ex- of real score change at various levels of initial test scores.

Analysis of Score Change

The procedure employed here in estimating score change follows a procedure described by F.M. Lord (personal communication, August 19, 1980). It depends on accessing and ex-

pressing scores for equating sections of the SAT on the same scale as the regular verbal or mathematical scores reported to test candidates. Let X denote a regular SAT-V or SAT-M score and Y a score on a corresponding verbal or mathematical equating section. All X scores on different forms are equated on a common scale (i.e., the 200-800 point scale for the SAT), and Y scores on different sections can also be expressed on a common scale. Further assume that there is a linear relationship between true scores on X and Y. Score X contains a true score, T, and an error of measurement, E:

$$X = T + E.$$

Score Y similarly contains a true score, aT + b where a and b are scaling constants, and an error, F:

$$Y = aT + b + F.$$

After taking X_1 and Y_1 for the first time as juniors, a subgroup of examinees decides to retake X_2 and Y_2 as seniors. Another random sample of examinees has taken X_0 and Y_0, here chosen from the same initial test administration in the junior year for convenience. It is assumed that all errors are unbiased with an expected value of zero except E_1, because the decision to retake the SAT was partially based on score X_1. Averaging across examinees it will be found that:

$$\bar{X}_0 = \bar{T}_0, \qquad \bar{Y}_0 = a\bar{T}_0 + b;$$
$$\bar{X}_1 = \bar{T}_1 + \bar{E}_1, \quad \bar{Y}_1 = a\bar{T}_1 + b;$$
$$\bar{X}_2 = \bar{T}_2, \qquad \bar{Y}_2 = a\bar{T}_2 + b.$$

From this set of relationships we can determine the scaling constants as well as $\bar{T}_1$:

$$a = (\bar{Y}_2 - \bar{Y}_0)/(\bar{X}_2 - \bar{X}_0),$$
$$b = \bar{Y}_0 - a\bar{X}_0 = \bar{Y}_2 - a\bar{X}_2,$$
$$\bar{T}_1 = (\bar{Y}_1 - b)/\, a.$$

The estimate of average true score change among all candidates repeating the test as X_1 and X_2 is then:

$$\bar{T}_2 - \bar{T}_1 = \bar{X}_2 - \frac{1}{a}\ (Y_1 - b) = \frac{\bar{X}_2 - \bar{X}_0}{\bar{Y}_2 - \bar{Y}_0}\ (\bar{Y}_2 - \bar{Y}_1).$$

This estimate of average true score change from examinee performance on equating sections and scaling constants can be contrasted with the average observed score change, $\bar{X}_2 - \bar{X}_1$. If the negative errors of measurement on initial test scores encouraged some students to retake the test, it should be the case that $\bar{E}_1$ is nonzero and negative,

$$\bar{E}_1 = \bar{X}_1 - \bar{T}_1 \text{ with } \bar{X}_1 - \bar{T}_1 < 0,$$

and that

$$\bar{X}_2 - \bar{X}_1 > \bar{T}_2 - \bar{T}_1.$$

For individuals an estimate of true score change based on their performance on equating sections would be:

$$T_2 - T_1 = \frac{\bar{X}_2 - \bar{X}_0}{\bar{Y}_2 - \bar{Y}_0} (Y_2 - Y_1) + F_1 - F_2.$$

This $T_2 - T_1$ may be averaged over individuals at the same observed initial score level, in order to show how $T_2 - T_1$ varies with X_1, since $F_1 - F_2$ should average to zero. These estimates of $T_2 - T_1$ at different initial score levels should result in a slope along X_1 less steep than that for $X_2 - X_1$ along X_1 as the latter observed score difference will be subject to regression effects while the former estimated score difference should be independent of errors of measurement on the initial test.

Samples of Test Candidates

Samples of test records representative of the most common pattern of test repetition among secondary school students were drawn from data files. These records came from the administrations of the SAT given in May 1979 and November 1979, and included only students who had taken the SAT for the first time in May 1979 as juniors and repeated the test in November 1979 as seniors without any intervening administrations of the test. There were 81,959 such students, but the analysis of score change also required scores on verbal or mathematical equating sections for both test administrations. Of the ten experimental sections spiraled in the administration of the SAT given in May 1979, there were two verbal and two mathematical equating sections. Of the ten experimental sections spiraled in the administration of the SAT given in November 1979, there were three verbal and three mathematical equating sections. Scores on experimental sections were available for all 31,971 students who had taken either a verbal or mathematical equating section in May 1979 and fit the May-November pattern for test repetition. It was possible to retrieve only 129,878 records with scores on experimental sections from the test administration given in November 1979 and just 9,148 records could be matched against the 31,971 cases from the test administration of May 1979. A total of 5,277 examinees had taken equating sections in both test administrations and were matched through these data files from May 1979 and from November 1979. Of these there were 1,325 examinees with scores on verbal equating sections for both test administrations (Y_1 and Y_2 on verbal equating sections corresponding to X_1 and X_2 on regular verbal sections) and 1,312 examinees with scores on mathematical equating sections for both test administrations (Y_1 and Y_2 on mathematical equating sections corresponding to X_1 and X_2 on regular mathematical sections).

It was also necessary to convert scores on equating sections to a common scale as a prerequisite to the proposed analysis of score change. This could be accomplished in a straightforward manner within each test administration since the spiraling of experimental sections among test candidates should result in comparable groups of examinees taking each section and thereby permit the linking of scores based on means and standard deviations. A verbal equating section from the test administration given in May 1979 and another from the test administration given in November 1979 were among the experimental sections spiraled in the administration of the SAT in November 1980, which established a link between the equating sections from May 1979 and from November 1979. The same spiraling occurred in November 1980 for a mathematical equating section from the test administration given in May 1979 and for another mathematical equating section from the test administration given in November 1979. Spiraling of experimental sections made it possible to link equating scores within and between test administrations.

Because the data files with results on experimental sections from the test administration given in November 1980 were incomplete and because the matches of these data files

from May and November 1979 depended on linking equating scores within and between test administrations, a second sample of the same pattern of test repetition was drawn using complete data files from the test administration of May 1979 and random samples of equating sections kept from the test administration of November 1979. This sample was restricted to 1,020 examinees who had taken the particular verbal or mathematical equating sections which were included in the test administration of May 1979 and November 1979 and were subsequently spiraled together in the test administration of November 1980. Among these students there were 152 examinees with the respective pair of verbal equating sections (Y'_1 and Y'_2 on verbal equating sections corresponding to X'_1 and X'_2 on regular verbal sections) and 113 examinees with the respective pair of mathematical equating sections (Y'_1 and Y'_2 on mathematical equating sections corresponding to X'_1 and X'_2 on regular mathematical sections) for May 1979 and November 1979.

Random samples of test candidates who had taken the verbal equating section in May 1979, which was later also given in November 1980, as well as test candidates who had taken the corresponding mathematical equating section were available from data files routinely retained for the SAT (X_0 and Y_0).

RESULTS AND DISCUSSION

A summary of examinees' scores on reporting and equating sections of the administrations of the SAT given in May 1979 and November 1979 appears in Table 1. The random verbal and mathematical equating samples from the test administration given in May 1979, 1,800 and 1,755 test candidates respectively, provide information on test performance, X_0 and Y_0, necessary for the analysis of score change. The random verbal and mathematical equating samples drawn from students taking the SAT as juniors in May 1979, X'_1 and Y'_1, and repeating the test as seniors in November 1979, X'_2 and Y'_2, represent a merger of particular pairs of equating sections based on complete records for test candidates from May 1979 and on random files of test candidates from November 1979; these samples of 152 and 113 test candidates with verbal or mathematical equating sections in both test administrations depended on only one link between equating sections in forming a common score scale. The largest samples of test candidates with results on reporting and equating sections as juniors and seniors, 1,325 students with verbal equating sections in both test administrations (i.e., Y_1 and Y_2 in the verbal equating sample for file matches) and 1,312 students with mathematical equating sections in both test administrations (i.e., Y_1 and Y_2 in the mathematical equating sample for file matches), represent a match of complete data files from May 1979 and incomplete data files from November 1979; these samples depended on multiple links within and between test administrations in forming a common score scale for equating sections.

The test results in Table 1 show that secondary school juniors electing to repeat the SAT as seniors do almost as well on their initial test administration as their peers but as seniors do noticeably better on their second test administration than their peers. Only two points separate the SAT-Verbal scores and only three points the SAT-Mathematical scores of all junior examinees and junior-senior repeaters on the test administration for the junior year. But the junior-senior repeaters exceed the performance of all senior examinees by 16 points on SAT-Verbal scores and 22 points on SAT-Mathematical scores in the senior year. These latter differences probably reflect as much on idiosyncracies in student choices of test administrations and the ability of the respective groups as on the score change possible through practice and growth, especially since junior examinees overall attained higher SAT-V and SAT-M scores in May 1979 than did senior examinees from the same high school cohort in November 1979.

TABLE 1. Means and Standard Deviations of Reporting and Equating Scores Across Test Administrations

Group	N	Junior Year (May 1979) SAT-Verbal Mean	SAT-Verbal sd	SAT-Math. Mean	SAT-Math. sd	Equating Sect. Mean	Equating Sect. sd	Senior Year (November 1979) SAT-Verbal Mean	SAT-Verbal sd	SAT-Math. Mean	SAT-Math. sd	Equating Mean	Equating Sect. sd
All Examinees	253,354	432	107	478	113								
Junior examinees	223,394	439	105	486	111								
Verbal equating sample (X_0, Y_0)	1,800	429.76	107.33			16.26	7.94						
Mathematical equating sample (X_0, Y_0)	1,755			481.75	114.81	10.25	6.06						
All Examinees	348,954							434	107	476	114		
Senior examinees	329,601							434	106	477	114		
Junior-Senior Repeaters	81,959	437	97	483	104			450	99	498	105		
Random sample	1,020	435.67	97.27	481.36	102.36			449.44	98.43	493.19	103.41		
Verbal equating sample (X_1', Y_1'; X_2', Y_2')	152	427.43	93.97	484.41	107.01	16.30	7.29	447.63	94.64	490.86	111.97	17.85	7.15
Mathematical equating sample (X_1', Y_1'; X_2', Y_2')	113	437.52	99.38	477.43	89.91	10.35	5.17	452.39	105.23	498.85	98.15	11.38	5.71
File matches	5,277	436.85	99.89	480.45	106.20			449.55	99.94	496.58	106.31		
Verbal equating sample (X_1, Y_1; X_2, Y_2)	1,325	439.93	98.24	483.32	104.21	16.87	7.43	452.42	98.21	498.56	104.95	18.63	7.76
Mathematical equating sample (X_1, Y_1; X_2, Y_2)	1,312	435.99	99.49	478.45	106.58	10.24	5.69	448.32	99.25	495.55	105.79	10.84	5.68

TABLE 2. Intercorrelations Among Reporting and Equating Scores in the Junior and Senior Year

		Junior Year			Senior Year		
		SAT-V	SAT-M	Equating Section	SAT-V	SAT-M	Equating Section
Junior year	SAT-V		.63	.86	.88	.62	.83
	SAT-M	.61		.60	.60	.87	.60
	Equating section	.58	.85		.86	.59	.79
Senior year	SAT-V	.88	.58	.56		.60	.85
	SAT-M	.59	.87	.83	.57		.60
	Equating section	.55	.82	.79	.54	.85	

Note: Entries above the diagonal reflect correlations for the 1,325 repeaters with a verbal equating section in both their junior and senior year; entries below the diagonal reflect correlations for the 1,312 repeaters with a mathematical equating section in both their junior and senior year.

Table 2 presents correlations between reporting and equating scores for students taking the SAT as juniors and seniors. The high correlations between SAT-Verbal scores across years, $r_{X_1X_2} = .88$, and between SAT-Mathmatical scores across years, $r_{X_1X_2} = .87$, imply a high degree of consistency in test performance from the junior year to the senior year. Similar stability is reflected in correlations between corresponding reporting and equating sections within years, $r_{X_1Y_1}$ and $r_{X_2Y_2} = .85$. These findings would seem to weaken the argument that students decide to repeat the test in part because they perceive that negative errors of measurement have depressed their initial test scores. Such errors should also lessen the correlation of test scores across occasions unless errors of measurement persist across occasions for certain individuals (i.e., $\rho_{E_1E_2} > 0$). Table 2 also shows that the correlation between SAT-Verbal and SAT-Mathematical scores remains low among students repeating the SAT as juniors and seniors, $r = .60$, in contrast to the correlation found among students taking the SAT only once as juniors, $r = .73$ (Alderman, 1981). This both strengthens the possibility that errors of measurement persist across occasions and suggests an alternative explanation for some students deciding to retake the test, as they perhaps believe that a disparity between verbal and mathemathical scores signifies an opportunity for improving the lower score upon test repetition rather than relative strengths in the respective abilities.

Score changes appear in Table 3 as observed differences between test administrations in the junior and senior years, $\bar{X}_2 - \bar{X}_1$, and estimated differences between test administrations in the junior and senior years, $\bar{T}_2 - \bar{T}_1$, along with the average error of measurement on the initial test administration, $\bar{E}_1$. For the random equating samples the expected relationships hold between observed score changes and estimated score change and tend to confirm student self-selection in deciding to repeat the test on the basis of negative errors of measurement on the initial test administration (i.e., $\bar{X}_2 - \bar{X}_1 > \bar{T}_2 - \bar{T}_1$ and $\bar{E}_1 < 0$). However, it should also be noted that the largest average verbal score gain and the largest average mathematical score gain, as reported above in Table 1, occurred in these respective random equating samples.

The larger samples drawn from matches of data files from the pertinent test administrations offer conflicting evidence about the necessity for this procedure for the

TABLE 3. Observed and Estimated Scores Changes from Junior to Senior Year

Group	N	Observed Mean Change	Estimated Mean Change	Estimated Measurement Error
SAT-Verbal				
Random samples $(X'_1, Y'_1; X'_2, Y'_2)$	152	20.20	17.48	-2.72
File matches $(X_1, Y_1; X_2, Y_2)$	1,325	12.49	16.84	+4.35
SAT-Mathematical				
Random Samples $(X'_1, Y'_1; X'_2, Y'_2)$	113	21.42	15.53	-5.88
File matches $(X_1, Y_1; X_2, Y_2)$	1,312	17.10	14.03	-3.07

analysis of score change and about negative errors of measurement on initial scores prompting students to repeat a test. Although the observed score change exceeds the estimated score change for the SAT-M, the reverse is the case for the SAT-V, as shown in Table 3. Yet a positive error of measurement should lead students to accept their initial test scores rather than to repeat the test. Of course this estimated positive error of measurement on initial test scores may simply reflect the large standard error of measurement inherent to difference scores or the inevitable error arising from the complexity of links among equating sections within and between test administrations.

Although the particular procedure for the analysis of score change followed here does yield estimates of average score change closer together than observed score differences for the random samples and the file matches (i.e., 17.48 and 16.84 versus 20.20 and 12.49 for verbal scores, and 15.53 and 14.03 versus 21.42 and 17.10 for mathematical scores), the procedure may not fully compensate for student self-selection in test repetition. Adjustments for student self-selection based on equating scores alone risk a correlation between errors of measurement on equating and reporting sections (i.e., $\rho_{E_1F_1} > 0$). For example, atypical test performance attributable to a student's physical health or emotional stress is apt to affect scores on all sections. Moreover, student self-selection in deciding to repeat a test may occur because scores seem inconsistent with high school rank, years of mathematics study, years of English study, and other variables aside from scores on equating sections (see Alderman, 1981).

Figures 1 and 2 illustrate the appropriateness and usefulness of the estimates of score change in adjusting for the effects of regression when students repeat tests. The points plotted in these figures represent observed and estimated average true score change by level of initial test score for the samples formed by matching data files (see Appendix A). The solid lines represent the best least-squares linear fit for observed score changes, and the broken lines the same for estimated score changes. The negative slopes for observed score change by initial test score clearly reflect

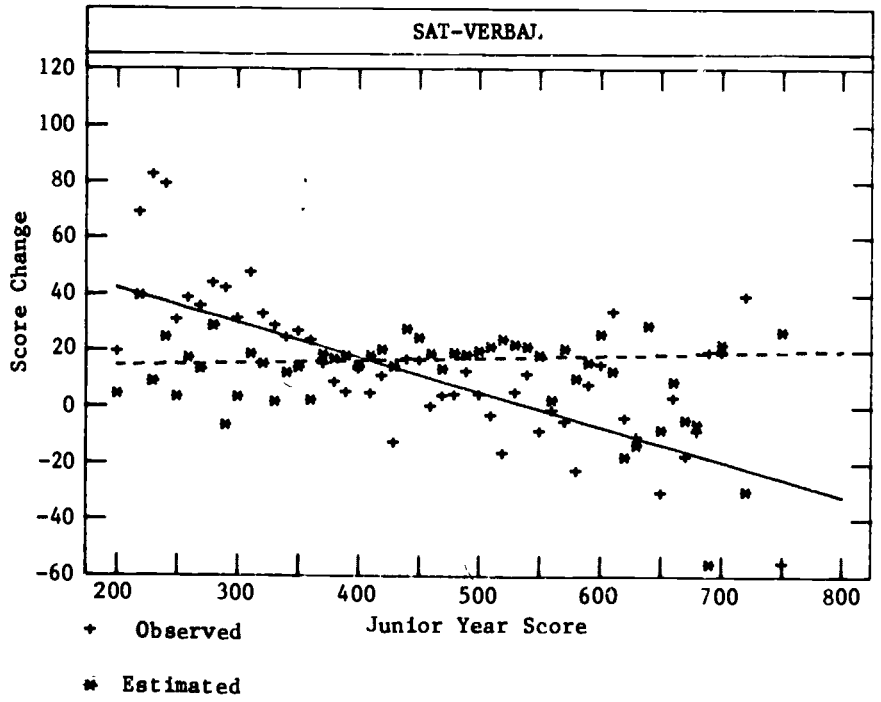

FIGURE 1. Observed and Estimated Score Change by Initial Level of SAT-Verbal Scores.

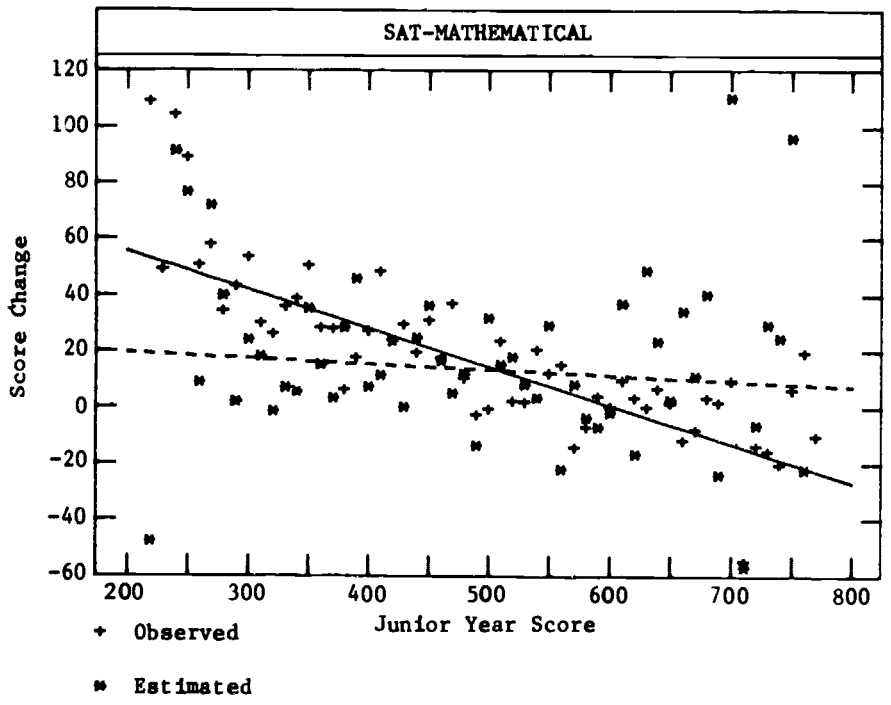

FIGURE 2. Observed and Estimated Score Change by Initial Level of SAT-Mathematical Scores. Points for estimated score change at initial score levels of 230 and 770 fall beyond the bounds of the abscissa.

the expected negative correlations indicative of scores regressing toward the mean upon retesting, $r_{(X_2-X_1)X_1} = -.249$ for verbal scores and $r_{(X_2-X_1)X_1} = -.269$ for mathematical scores. These correlations between observed score change and initial test score correspond closely to their respective expected values (i.e., -.232 for verbal scores and -.259 for mathematical scores). The slopes of nearly zero for estimated true score change by initial test score, $r_{(T_2-T_1)X_1} = -.020$ for verbal scores and $r_{(T_2-T_1)X_1} = -.025$ for mathematical scores, contrast sharply with those for observed score change by initial test score and show that true score change is nearly constant across different levels of initial test scores. The extent to which measurement error affects score change as initially low scores increase and initially high scores decrease is shown by the difference in slopes between observed score change in these figures.

Rescaling scores on equating sections to the same scale as reported for regular SAT scores results in estimates of score change seemingly free from the typical regression effect arising from measurement error. This suggests the independence of errors of measurement on equating and reporting sections, an implicit assumption for this analysis of score change. Further, it would appear to be the case that the errors of measurement inherent to SAT-Verbal and SAT-Mathematical scores stem primarily from variation in content, difficulty, and other aspects of the specifications for parallel test forms and from variation in individual responses like guessing, pacing, and omitting across test sections. Sources of measurement error like an individual's health or the conditions of test administration would likely affect equating scores as well as reporting scores, and thereby lead to parallel slopes for observed and estimated score change by initial test score rather than the marked differences in slopes apparent in Figures 1 and 2. Chronic poor test performance may prompt some students to retake the SAT if such a component of measurement error remains fairly constant across test sections and administrations and is independent of other sources of measurement error. The majority of secondary school juniors repeating the test probably do so because they already planned test repetition as seniors or perceived their scores on the initial test administration as underestimates of their abilities.

While the use of equating sections or other types of parallel and concurrent test forms might yield better estimates of score change across levels of initial test scores, such a procedure is dependent upon aggregating data at those score levels and would still give an unreliable and unstable estimate of score change for an individual. Errors of measurement on equating sections (i.e., F_1 and F_2) are a part of the estimated score change for individuals. The standard deviations for the observed and estimated verbal score changes depicted as averages in Figure 1 were 48.86 points and 46.70 points, respectively. For the observed and estimated mathematical score changes in Figure 2 the standard deviations were 54.40 points and 85.57 points, respectively. Although the standard deviations for observed and estimated verbal score changes are comparable, the standard deviation for estimated mathematical score changes is obviously greater than the standard deviation for observed mathematical score changes. This simply reflects the greater opportunity for errors of measurement in sampling mathematical ability on an equating section which is half the length of the regular sections entering into the scores actually reported for test candidates. These standard deviations indicate that neither observed nor estimated score change accurately represents a particular individual's change in abilities between test administrations.

SUMMARY AND CONCLUSIONS

Despite the fact that the average score change on the Scholastic Aptitude Test is approximately one-half of the test's standard error of measurement, the majority of students

electing to repeat the SAT actually experience score changes greater than the standard error of measurement. These gains and losses between test administrations largely reflect the influence of measurement errors as students with low initial scores stand to gain points upon retesting while students with high initial scores may lose points. Score changes on admissions tests such as the SAT, therefore, represent unreliable measures of individual growth and development in the abilities tapped by the test.

A procedure for estimating true score change resulted in a nearly constant estimate of score change across levels of initial SAT-Verbal and SAT-Mathematical scores. This procedure took the regression effect on score change and possible student self-selection in test repetition into account through the use of equating sections, essentially parallel and concurrent shorter forms of the SAT-V and SAT-M which do not affect reported scores, in obtaining independent estimates of score change between test administrations. The success of the procedure in yielding estimates of score change apparently free from the effects of measurement error further suggests that the errors of measurement leading to the regression effect on score change arise primarily from differences in content, difficulty, and other aspects of test specifications for parallel forms and from differences in guessing, pacing, omitting, and other components of individual responses across test sections. Other procedures, however, would probably be as successful in adjusting for these same sources of measurement error (e.g., Lord, 1963).

The evidence was equivocal regarding possible student self-selection in deciding to repeat an admissions test because they perceive their initial test scores as underestimates of their actual abilities. For some samples of test candidates there were the expected estimates of overall negative errors of measurement on initial test scores. But there was an estimate of an overall positive error of measurement for the initial verbal scores of a large number of secondary school juniors later repeating the test as seniors. This may simply reflect the error involved in linking the score scales of various equating sections or indicate the importance of background variables like high school rank and years of study in different subject matters in student perceptions of the consistency of their test scores with their academic abilities--factors which the analysis of score change did not take into account. Lower correlations between verbal and mathematical scores among students with repeat test administrations than among students with a single test administration also suggest that some students may retake an admissions test because they believe differences in scores for separate traits to be a discrepancy rather than an indication of their relative strengths. Nevertheless, the analysis of score change resulted in estimated differences between test administrations more consistent across levels of initial test scores and across samples of test candidates than were the observed score differences.

The procedure followed here in estimating true score change depends on scores from concurrent and parallel tests on two occasions. It succeeds in correcting for regression effects when the correlation between parallel test forms administered at the same time (i.e., $r_{X_1Y_y}$ and $r_{X_2Y_2}$) is nearly the same as the correlation between separate test administrations (i.e., $r_{X_1X_2}$). Under these conditions the regression effects noted in examining observed score changes by initial levels of test performance apparently arises from errors of measurement also found across parallel forms, and the procedure accounts for such variations in aspects of test specifications and in responses of test candidates. Under other conditions, for example when the correlation between parallel forms is much greater than the test-retest correlation, the same procedure would probably not account for all of the sources of variation leading to regression effects.

Although it has been common practice to relate average score changes on admissions tests to their standard error of measurement, such a comparison seems inappropriate when change involves two test administrations and when change in the relevant abilities might occur over the time between test administrations. Difference scores would then be susceptible to errors of measurement on both the initial and a subsequent test administration. Indeed, the majority of examinees repeating the SAT experience score changes greater than the standard error of measurement. A more appropriate benchmark for

judging the meaningfulness of score change might be the standard error of estimate coupled with the intercept constant in predicting subsequent scores from initial scores. Regardless of the basis for comparison, the inherent unreliability of difference scores suggests that little weight can be given to the observed change in admissions or placement decisions.

REFERENCES

Alderman, D. L. "Student Self-Selection and Test Repetition," Educational and Psychological Measurement, 1981, in press.

Donlon, T. F., and Angoff, W. H. "The Scholastic Aptitude Test." In The College Board Admissions Testing Program, W. H. Angoff (Ed.) Princeton, N.J.: College Entrance Examination Board, 1971, pp. 15-47.

Duggan, J. M. "Puzzles and Powers in Junior SAT Scores," College Board Review, 37 (1959): 37-39.

Educational Testing Service. ATP Guide for High Schools and Colleges 1979-81. New York: College Entrance Examination Board, 1979.

Educational Testing Service. An SAT: Test and Technical Data. New York: College Entrance Examination Board, 1980.

Jacobs, P. I. "Large Score Changes on the Scholastic Aptitude Test," Personnel and Guidance Journal, 45 (1966): 150-156.

Kendrick, S. A. "When SAT Scores Go Down," College Board Review, 64 (1967): 5-11.

Linn, R. L., and Slinde, J. A. "The Determination of the Significance of Change Between Pre- and Post-Testing Periods," Review of Educational Research, 47 (1977): 121-150.

Lord, F. M. "The Measurement of Growth," Educational and Psychological Measurement, 16 (1956): 421-437. [Errata, "The Measurement of Growth," Educational and Psychological Measurement, 17 (1957): 452.]

Lord, F. M. "Further Problems in the Measurement of Growth," Educational and Psychological Measurement, 18 (1958): 437-451.

Lord, F. M. "Elementary Models for Measuring Change," in Problems in Measuring Change, C. W. Harris (Ed.) Madison, Wisc.: University of Wisconsin Press, 1963, pp. 21-38.

McNemar, Q. "On Growth Measurement," Educational and Psychological Measurement, 18 (1958): 47-55.

Messick, S. The Effectiveness of Coaching for the SAT: Review and Reanalysis of Research from the Fifties to the FTC (ETS Research Report RR-80-8). Princeton, N.J.: Educational Testing Service, 1980.

Nesselroade, J. R., Stigler, S. M., and Baltes, P. B. "Regression Toward the Mean and the Study of Change," Psychological Bulletin, 88 (1980): 622-637.

Overall, J. E., and Woodward, J. A. "Unreliability of Difference Scores: A Paradox for Measurement of Change," Psychological Bulletin, 82 (1975): 85-86.

Rock, D., and Werts, C. An Analysis of Time-Related Score Increments and/or Score Decrements for GRE Repeaters Across Ability and Sex Groups (GRE Board Research Report GREB No. 77-9R). Princeton, N.J.: Educational Testing Service, 1980.

APPENDIX: Observed and Estimated Score Changes from Junior to Senior Year

Junior Year Score	Verbal Score Change N	Verbal Observed Mean Change	Verbal Estimated Mean Change	Mathematical Score Change N	Mathematical Observed Mean Change	Mathematical Estimated Mean Change
200	5	20.00	5.13	0		
210	0			0		
220	4	70.00	40.11	1	110.00	-47.50
230	3	83.33	9.38	1	50.00	-146.22
240	3	80.00	25.41	2	105.00	91.95
250	11	31.82	3.78	2	90.00	77.18
260	10	39.00	17.79	7	51.43	9.32
270	20	36.50	13.81	8	58.75	72.78
280	11	44.55	29.49	12	35.00	40.33
290	25	42.80	-6.15	8	43.75	2.55
300	10	32.00	3.34	14	54.29	24.44
310	39	48.46	19.29	18	30.56	18.53
320	25	33.60	15.39	22	26.82	-1.11
330	36	29.44	2.27	20	36.50	7.56
340	26	25.38	12.73	43	39.30	6.04
350	58	27.59	14.66	28	51.07	35.99
360	29	24.14	2.74	29	28.97	15.25
370	23	15.22	18.97	32	28.75	3.44
380	64	9.53	17.58	30	7.00	29.09
390	39	5.90	18.75	26	18.46	46.56
400	79	14.18	15.03	55	27.82	7.52
410	28	5.36	18.59	32	49.06	11.96
420	68	11.62	20.77	37	24.59	32.80
430	26	-11.92	15.07	30	30.33	.49
440	70	17.57	28.20	48	20.00	25.26
450	37	17.30	24.82	36	31.67	37.03
460	79	.89	19.48	90	18.44	17.10
470	29	4.48	13.73	44	37.73	5.85
480	52	4.81	19.83	40	11.00	12.24
490	33	13.03	19.19	34	-2.06	-13.20
500	51	5.10	20.43	36	0.28	32.37
510	26	-2.31	22.15	25	24.40	15.65
520	24	-15.83	24.49	71	2.68	18.63
530	49	5.71	22.64	40	2.50	8.59
540	16	12.50	22.06	33	21.21	3.37
550	40	-8.25	19.24	26	12.69	29.76
560	16	-.62	3.09	29	15.52	-21.66
570	42	-4.52	21.41	32	-13.75	8.56
580	15	-22.00	11.00	54	-6.67	-3.04
590	27	8.52	16.62	38	4.47	-6.34
600	18	15.56	26.48	13	.77	-1.46
610	9	34.44	12.96	16	10.00	37.67
620	3	-3.33	-17.10	20	4.00	-16.18
630	13	-10.00	-12.97	16	0.62	49.32
640	6	0.00	29.72	29	7.24	24.42
650	1	-30.00	-7.74	15	2.00	3.17
660	6	3.33	9.38	16	-11.25	34.98
670	10	-17.00	-4.21	13	-7.69	11.71
680	5	-8.00	-5.43	5	4.00	40.99
690	1	20.00	-55.91	4	2.50	-23.60

APPENDIX: Observed and Estimated Score Changes from Junior to Senior Year (continued)

Junior Year Score	Verbal Score Change N	Verbal Observed Mean Change	Verbal Estimated Mean Change	Mathematical Score Change N	Mathematical Observed Mean Change	Mathematical Estimated Mean Change
700	2	20.00	22.84	4	10.00	111.22
710	0			9	-54.44	-56.53
720	1	40.00	-29.45	6	-13.33	-5.63
730	0			4	-15.00	30.32
740	0			4	-20.00	25.23
750	2	-55.00	27.75	3	6.67	96.96
760	0			1	20.00	-21.82
770	0			1	-10.00	-64.55
780	0			0		
790	0			0		
800	0			0		
Total	1,325	12.49	16.84	1,312	17.10	14.03